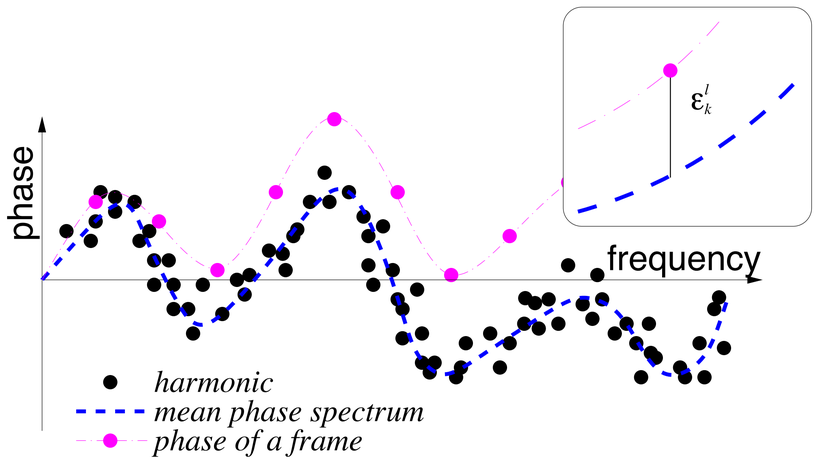
Estimating the Spectral Envelope of Voiced Speech Using Multi-frame Analysis

Estimating the envelope of phase spectrum



We first find the mean phase-spectrum, $\bar{\varphi}(f)$, of all the harmonics by taking a moving average of all the observed phases.

$$\bar{\varphi}(f) = \frac{\phi(f)}{|\phi(f)|}, \quad \phi(f) = \frac{\sum_{k=1}^{M}\sum_{l=1}^{N_k} v_k^l G(f_k^l - f) e^{j(\theta_k^l - 2\pi f_k^l \tau_k)}}{\sum_{k=1}^{M}\sum_{l=1}^{N_k} v_k^l G(f_k^l - f)}$$

θ_k^l : observed (wrapped) phases
τ_k : delay of frame k
G : moving average window
v_k^l : weight for harmonic l of frame k

The phase θ_k^l is then adjusted using the delay τ_k so that the following equation is minimised.

$$\sum_{k=1}^{M}\sum_{l=1}^{N_k} w_k^l \left\{\frac{\varepsilon_k^l}{2\pi f_k^l}\right\}^2 = \sum_{k=1}^{M}\sum_{l=1}^{N_k} w_k^l \left\{\frac{1}{2\pi f_k^l}\mathrm{ARG}\left[\frac{e^{j(\theta_k^l - 2\pi f_k^l \tau_k)}}{\bar{\varphi}(f_k^l)}\right]\right\}^2$$

We can obtain the correcting value $\Delta\tau_k$ for the delay τ_k as:

$$\Delta\tau_k = \sum_{l=1}^{N_k} \frac{w_k^l}{2\pi f_k^l}\mathrm{ARG}\left[\frac{e^{j(\theta_k^l - 2\pi f_k^l \tau_k)}}{\bar{\varphi}(f_k^l)}\right] \Big/ \sum_{l=1}^{N_k} w_k^l$$

4. Experiment

Experimental condition

- We used the following corpus with parallel acoustic-articulatory information.

MOCHA-TIMIT corpus

speaker		female (fsew0)
number of sentences		460
sampling rate	speech	16.0 kHz
	EMA data	500 Hz

- 87208 voiced frames were extracted from the corpus using the following analysis.

Harmonic estimation

method		weighted LSM (Stylianou,2001)
analysis window	type	Hanning
	width	20.0 ms
	spacing	8.0 ms

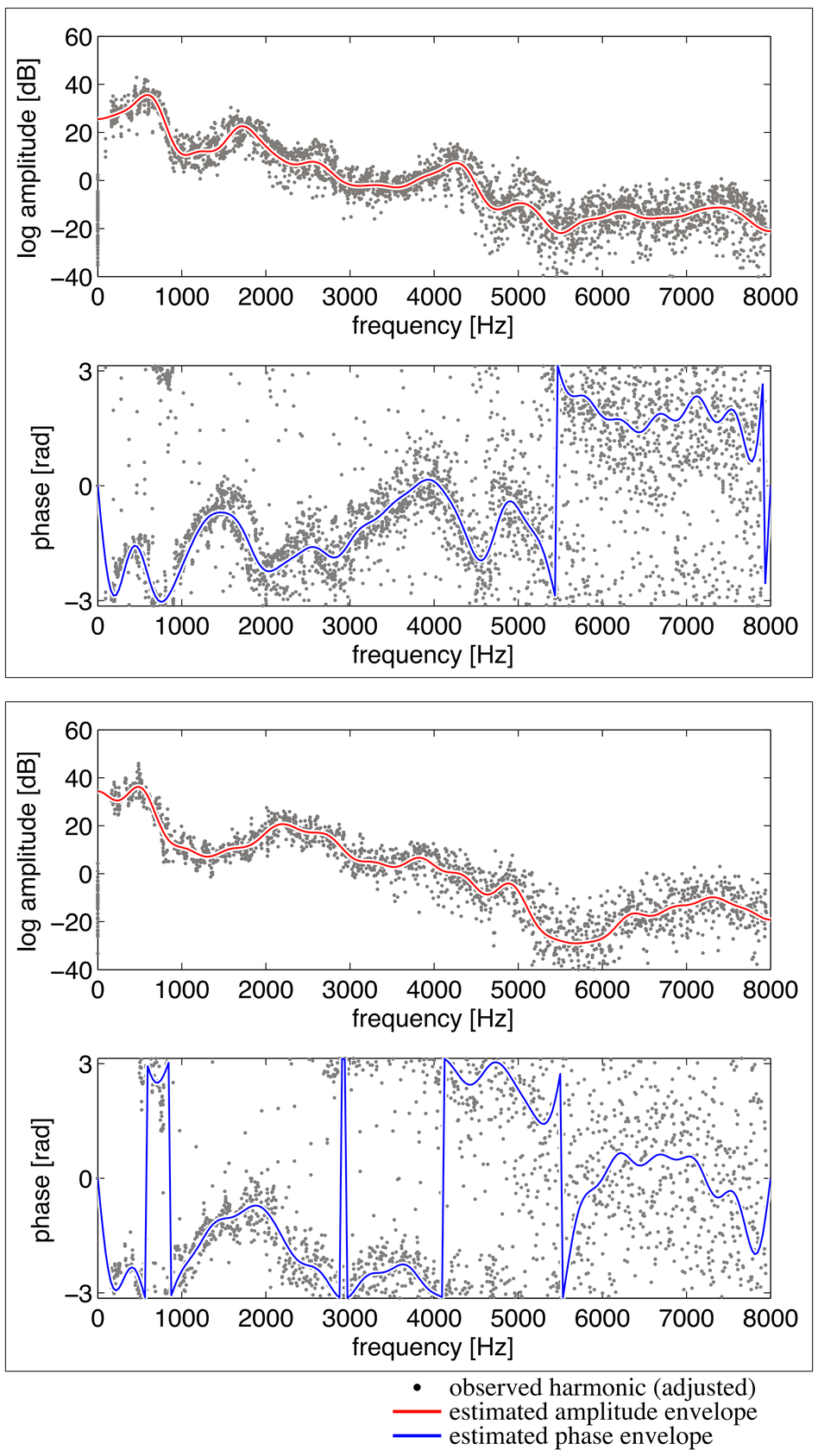
- **LBG clustering** (*Linde et al.,1980*) was applied to the articulatory data in order to identify frames with similar articulator settings. Cepstrum coefficients were then calculated by performing MFCA for all the frames in each cluster.

Multi-frame analysis

number of clusters	2048
order of cepstrum	40
w_k^l	$\exp\left[-\frac{(f_k^l)^2}{2(2[\mathrm{kHz}])^2}\right]/N_k$
v_k^l	$1/N_k$
$G(x)$	$\exp\left[-\frac{x^2}{2(100[\mathrm{Hz}])^2}\right]$

5. Results and discussion

Estimated envelopes



Discussion

- MFCA discovers smooth spectral **envelopes which best approximate** all the harmonics of all the frames within each articulatory cluster.
- Some clusters have comparatively large variances of observed harmonic spectra, which is probably because:
  - each frame has a different **noise level** (S/N) that changes depending on speech powers of the frames;
  - we do not take into account the variance in the **acoustic** space during the clustering in the **articulatory** space;
  - the voice source is **not a periodic impulse train** as we assumed, and its spectral characteristic changes depending mainly on F_0 and power.

6. Future work

What's next?

- In respect of the source problem, we have proposed an approach that can take into account the **change of the voice-source characteristic** using MFCA. ***(See Poster #)***
- The combination of MFCA and the articulatory clustering produces **a codebook which relates articulation with acoustic feature of speech**. With this codebook we are currently examining *articulation-to-speech conversion*, in which speech can be modified in *articulatorily-meaningful ways*.

To be continued on Poster # ⟹

Estimating the Spectral Envelope of Voiced Speech Using Multi-frame Analysis

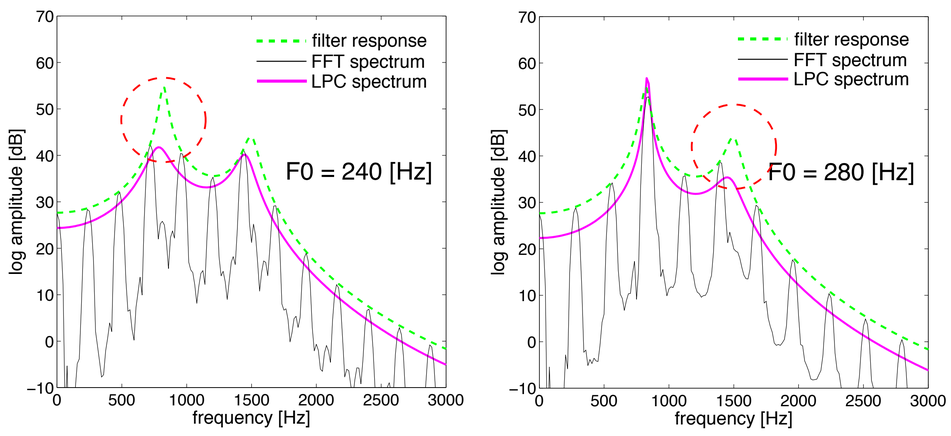
Yoshinori SHIGA and Simon KING
Centre for Speech Technology Research
University of Edinburgh
yoshi@cstr.ed.ac.uk

1. Motivation

Problem in spectral envelope estimation

The spectrum of voiced speech only has energy at frequencies corresponding to integral multiples of F_0, and therefore it is **impossible to identify the transfer characteristics between the harmonics**

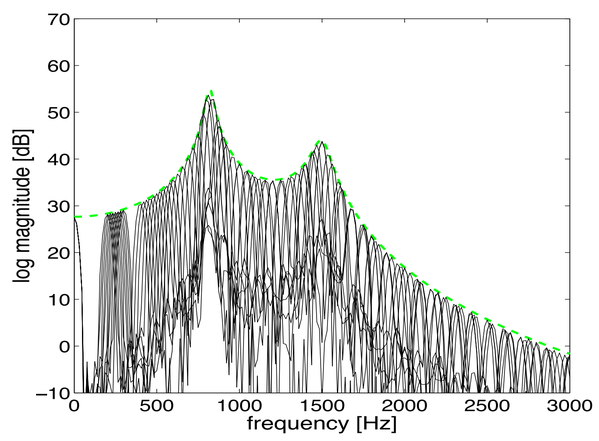
Conventional methods



Spectral envelope estimation is **interfered with by the harmonic structure** in conventional methods.

2. The idea

Using multiple speech-frames



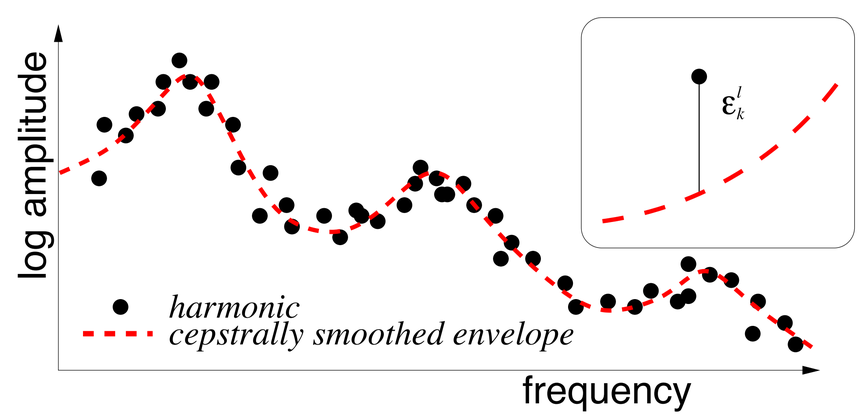
Using **multiple speech signals** generated with **different F_0** through the **same transfer function** allows us to estimate more exact envelopes.

3. Multi-frame analysis

Multi-frame cepstral analysis (MFCA)

We employ the **cepstrum** as an expression of the spectral envelope.

Estimating the envelope of amplitude spectrum



To find a cepstrum which best approximates the amplitude of all the harmonics in multiple frames we use **least squares estimation** in the frequency domain, which is an extension of the method by *Galas and Rodet (1990)*.

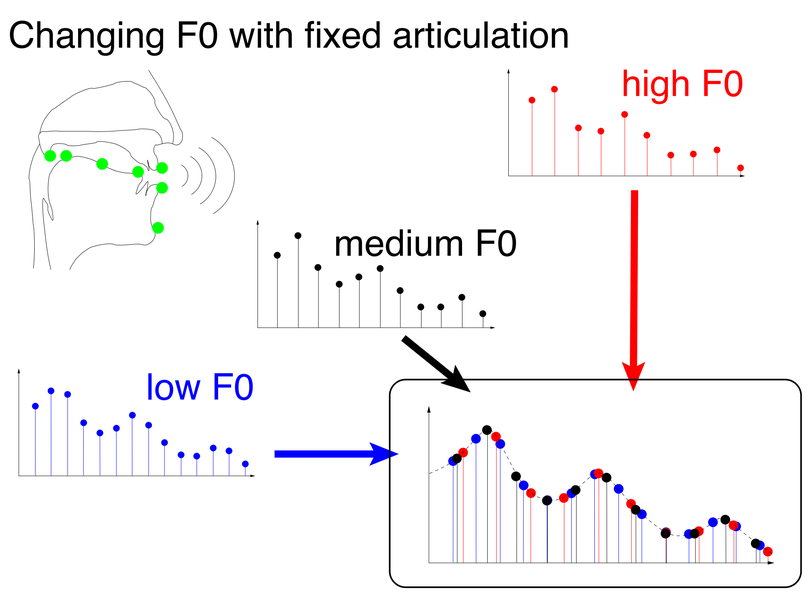
The sum of the squares of approximation error for all the harmonic amplitudes of all the frames is as follows:

$$\sum_{k=1}^{M}\sum_{l=1}^{N_k} w_k^l(\varepsilon_k^l)^2 = \sum_{k=1}^{M}\sum_{l=1}^{N_k} w_k^l \left[a_k^l - d_k - \sum_{n=-p}^{p} c_n \cos(2\pi f_k^l T n)\right]^2$$

a_k^l : observed log-amplitude of harmonic l in frame k
f_k^l : observed frequency of harmonic l in frame k
c_n : cepstral coefficients
N_k : number of harmonics in frame k
M : number of frames
T : sampling period
w_k^l : weight for harmonic l in frame k
d_k : offset that adjusts the total power (c_0)

This can be solved by reducing it to a problem of weighted least squares.

How to collect the frames



- In a **large speech corpus**, there must exist multiple speech portions produced using *same vocal-tract shapes*.
- Those portions usually have **different F_0s**.
- **Electro-magnetic articulograph (EMA)** data indicates where identical vocal-tract shapes are in the corpus.
- Otherwise, phonetic information, such as phonemic context, could be used to identify those locations.